Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 292

Introducing Social Norms in Game Theory

Raúl López-Pérez

June 2006

Introducing Social Norms in Game Theory*

Raúl López-Pérez†

June 2006

Abstract

This paper explicitly introduces norms in games, assuming that they shape (some) players' utility and beliefs. People feel badly when they deviate from a binding norm, and the less other players deviate, the more badly they feel. Further, people anger at transgressors and get pleasure from punishing them. I then study how social norms and emotions affect cooperation, coordination, and punishment in a variety of games. The model is consistent with abundant experimental evidence that alternative models of social preferences cannot explain.

Keywords: Cooperation, Emotions, Focal Points, Punishment, Reciprocity, Social Norms. JEL classification numbers: C72, D02, D62, D64, Z13.

---

* I am indebted to Ernst Fehr, Urs Fischbacher, Alexander Kritikos, Michael Kosfeld, Michael Näf, Christian Zehnder, participants at the March 2006 Greifensee seminar, and at the May 2006 Conference in Capua (Italy) for helpful comments. Part of this research was conducted while visiting the Institute for Empirical Research in Economics at Zurich, and I would like to thank their members for their hospitality. I also gratefully acknowledge financial support from the European Union through the ENABLE Marie Curie Research Training Network.

† Institute for Empirical Research in Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland. E-mail: rlopez@idea.uab.es

1. Introduction

A *norm* is a rule that prescribes behavior –that is, any statement of the form 'in situation *x*, you *ought to* do *y*'. For instance, all laws, codes of honor, moral principles, or religious commandments are norms according to this wide-ranging definition.

Prominent social researchers like Emile Durkheim or Talcott Parsons have pointed out the vital role that norms play in the attainment of social order –see also Arrow (1974), and Elster (1989). According to this view, most social norms commend pro-social behavior and hence are crucial to promote cooperation and social cohesion, which are basic ingredients for a society to exist.

Of course, this begs the question of *why (and when) people respect norms*. This paper offers a simple game-theoretical model to address such a question. The model applies on any game of perfect recall and keeps the standard assumptions that all players are rational and able to form accurate expectations about other players' behavior. However, it relaxes the standard selfishness hypothesis that *all* players are *exclusively* motivated by their *own* material interest –that is, their own consumption and leisure.

Instead, the model assumes that people *also* care about norms, which means two things. First, and in line with Classical Sociology (Parsons, 1967) and Social Psychology, *norms shape preferences*. When someone internalizes a norm (Elster, 1989; Becker, 1996; Gintis, 2003), she becomes emotionally attached to it so that painful *emotions* get triggered when she transgresses the norm (shame, guilt), or when others deviate (anger, indignation).[1] Second, *norms also shape beliefs* by acting as focal points in games (Schelling, 1960; Sugden, 1989). More precisely, players find obvious or prominent an equilibrium in which players respect internalized norms -if such equilibrium exists-, and that coordinates beliefs about co-players' behavior.

So, when do people obey norms? As agents care about both norms and material interest, the interesting case appears when respect for an *internalized* norm is at odds with one's material interest. The model then predicts compliance either if the expected remorse (*internal punishment*) is intense enough or if the probability of being heavily sanctioned by others (*external punishment*) is high enough. Further, and because I also assume that a deviator's bad feelings intensify if most others comply, people respect internalized norms in a *reciprocal* manner, that is, they are more willing to comply if others comply as well.

Apart of exploring why people respect norms, this paper also studies a key empirical question, that is, *what actual social norms are like*.[2] I focus on norms of *distributive justice* and show that if (some) agents have internalized a norm exhibiting a

[1] Further, pleasant emotions like pride or admiration can be activated if one or others, respectively, respect an internalized norm.

[2] Roughly, a norm is social if a 'large' proportion of the population has internalized it.

concern for both efficiency *and* maximin (the *EM-norm*)[3] the model then explains a large and varied array of well-replicated experimental results. This rough test suggests that such norm (or a similar one) is social –I have studied alternative norms in López-Pérez (2004).

This will be of great interest for experimental economists, who have gathered in the last 30 years an impressive amount of evidence contradicting the standard selfishness hypothesis. The model explains, for instance, why people cooperate conditionally, why moving first in a sequential social dilemma makes people relatively more cooperative than in a simultaneous one, why punishment and cooperation depend on the menu of choices, why passive players are not punished, why (and when) social norms increase coordination, or why competitive markets induce principled people to behave as self-interested ones do.

The model is related to recent theories of social preferences and reciprocity, which also relax the neoclassical hypothesis that all agents are selfish.[4] Rabin (1993) models reciprocity in normal-form games as the idea that people are kind to those who are kind to them, and unkind to those who are unkind. Dufwenberg and Kirchsteiger (2004) extend Rabin's ideas to extensive form games. Levine (1998) assumes type-based altruism and spitefulness, and both Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) propose models of inequity-averse players. Finally, Charness and Rabin (2002) and Falk and Fischbacher (2006) introduce both reciprocity and distributional concerns.

Although none of these models explicitly introduces norms,[5] they somehow admit an interpretation based on norms and hence provide some sensible intuitions on norm compliance. In spite of this, the model here has a number of advantages with respect to the other models.

First, it explains better the experimental evidence in the large range of games that I analyze here and in López-Pérez (2004). One crucial reason for this is that it assumes that agents care about *history*. More precisely, people's utility depends on whether others (or themselves) misbehaved –i.e., deviated from a binding norm- in the past. Hence, the model takes into account *procedural* justice. This is a key difference with outcome-based utility models like the inequity aversion ones.

Second, the model is relatively simple and precise. Contrary to most models of reciprocity, the model here is *not* based on the Psychological Game Theory of Geanakoplos et al. (1989), so that agents' utility does not depend on their beliefs. That makes the model much more parsimonious. Further, and contrary to Levine (1998), agents' utility does not depend on the co-players' types, and that prevents the existence of a multiplicity

[3] In this paper, efficiency refers to the sum of players' material payoffs, and not to Pareto efficiency. Maximin or need refers to the worst-off player's income.

[4] Fehr and Schmidt (2002), Camerer (2003), and López-Pérez (2004) survey this literature.

[5] Arguably, the only exception is the reciprocity model of Charness and Rabin (2002). This model is extremely complicated, though. Somehow, one might view my model as a tractable version of the ideas present in Charness and Rabin (2002).

of equilibria. In fact, my model predicts a unique equilibrium in most games, which is crucial to facilitate experimental testing.

Last, *but not least*, the model has a broader field of application because it explicitly introduces norms. One might use it to explain why people tell the truth and punish cheaters contrary to their material interest, or why people follow rules of etiquette, or norms regulating sexual relations. Other models have troubles in explaining such behavior because they posit that utility only depends on money allocations and/or on beliefs about allocations -and it is unclear how, say, sexual intercourse may affect those things!

The rest of the paper is organized as follows. Sections 2 and 3 describe the model and the efficient-cum-maximin (EM) norm, respectively. Section 4 studies how the EM-norm affects cooperation, coordination and punishment in different games, and shows the model to be consistent with abundant experimental evidence. The predictions of the model are briefly compared with those from other models in section 5. Section 6 concludes by mentioning possible extensions.

2. A Model with Social Norms

Consider a n-player, extensive form game of perfect recall Γ. Let N = {1,…,n} denote the set of players, z denote a terminal node, $u_i(z)$ denote player i's utility payoff at z, and $x_i(z)$ denote player i's *monetary* payoff at z.[6] The *lab game* associated to game Γ is obtained by substituting each monetary payoff $x_i(z)$ for its corresponding utility payoff $u_i(z)$ -distinguishing between lab games and proper games makes sense because, as I will posit later, $u_i(z)$ and $x_i(z)$ may differ for some players.

2.1 Norms

Norms are *exogenous* rules that select actions in *lab games* and indicate how human players *ought to* behave –e.g., 'Thou shalt not lie'. Let h denote an information set and $A(h)$ denote the set of available actions at h.

Definition 1: A norm is a nonempty correspondence $\Psi : h \to A(h)$ applying on any information set of any lab game, except on Nature's ones.

It is important to emphasize some ideas present in this definition. First, norms apply in lab games. Hence, no information about other player's utility payoffs is required to apply a norm, something important to avoid circularity problems. I will be more precise on this point later. Second, norms apply on *any* information set of *any* lab game. This may appear very demanding at first sight because many actual rules of proper behavior have a restricted field of application. For instance, the norm to wear black in funerals does not say anything about behavior at the workplace so one might be tempted to think that it is not a norm according to the previous definition. However, it is easy to accommodate such a

[6] More generally, $x_i(z)$ might be interpreted as a cardinal measure of the satisfaction that player i gets from consumption and leisure in the history of z.

norm by simply assuming that it commends *any* behavior at the information set 'workplace' –of course, other norms may be more restrictive.[7]

Given that norms select actions, a player is said to *respect* or *comply* with norm Ψ at h if (i) her choice at h is consistent with Ψ or if (ii) she does not move at h. Otherwise, she *deviates* from the norm. Suppose then that play reaches terminal node z. By considering all actions in the path of z, one may obtain the set of players who respected Ψ in the history of z, $R(\Psi, z)$, and its cardinality, $r(\Psi, z)$. If it is clear to which norm I refer, I will instead write $R(z)$ and $r(z)$.

2.2 Preferences

There are two types of players. S*elfish* players are standard money-maximizers who do not care about norms. Hence, their utility function is

$u_i(z) = x_i(z)$.

In contrast, the utility of a *principled* player at z depends on the money earned $x_i(z)$ *and* the history of z. In other words, principled agents care about *what* they get and *how* they get it. The intuition here is that different histories activate different emotions: If principled player A deviates from what an internalized norm commends then she feels ashamed or guilty, whereas if A complies but another player deviates then A feels angry at him.[8] More precisely, the utility function of a principled player i who has internalized norm Ψ equals

$$u_i(z) = \begin{cases} x_i(z) - \gamma \cdot r(z) & \text{if } i \notin R(z),\ (0 < \gamma) \\ x_i(z) - \alpha \cdot \max\limits_{j \notin R(z)} \{x_j(z)\} & \text{if } i \in R(z),\ (0 < \alpha \leq 1), \end{cases}$$

and it is convened that $\max\limits_{j \notin R(z)} \{x_j(z)\} = 0$ if nobody deviates –i.e., $R(z) = N$.

Parameter γ may be interpreted as a player's internalization index. Note that ceteris paribus the intensity of a deviator's bad feelings is assumed not to depend on the specific deviation she makes. That is, all deviations are equally 'bad'. Although this assumption is clearly unrealistic, it greatly simplifies the model and suffices to explain many experimental facts. I come back to this issue in the conclusion.

The more the people who respect the norm, the more badly a deviator feels. For simplicity, I have modeled this by means of a linear function, but any strictly increasing one would give the same qualitative results in the games I analyze. However, it is

[7] Throughout the paper, the following statements are synonymous: 'The norm selects action a at information set h', 'the norm commends to choose a at h', and 'according to the norm, (the relevant mover) should choose a at h.'

[8] Therefore, norms shape utility. This explains why norms are defined to apply on lab games and not on proper games: Circularity problems would appear if norms depended on utility payoffs (an ingredient of games) and at the same time affected utility.

important for the results that no principled *deviator* feels shame at z if *all* the other players deviate as well. As a result, a principled player *never* complies with an internalized norm if (a) compliance is at odds with her material interest and if (b) she *expects* all other player to deviate. This reciprocity idea will extensively appear in the applications.

Finally, parameter α measures aggressiveness –more precisely, an angry player i is willing to spend α monetary units in order to reduce the best-off deviator's monetary payoff in one unit. Note that α is independent of the specific deviation that triggers the anger, a hypothesis that is again made for simplicity –in the conclusion I discuss this issue. For analogous reasons, I also assume that anger and the associated tendency impulse to retaliate focus on the best-off deviator if there are multiple deviators.

2.3 Equilibrium Concept

Since I consider both simultaneous and sequential games, Subgame Perfect Equilibrium (SPE) is a natural solution concept to use. In addition, I introduce a refinement to model the idea that norms act as focal points. Let s', s'' denote a player's pair of *mixed* strategies of a game.

Definition 2: Strategy s' Ψ-dominates strategy s'' if s'' specifies with some positive probability a deviation from norm Ψ at an information set h where s' does not, and the opposite is not true at any other information set.

In other words, strategy s' Ψ-dominates strategy s'' if the corresponding player respects norm Ψ *unequivocally more* under strategy s' than under strategy s''. This domination concept need not provide a *complete* ordering of a player's strategies –i.e., it is logically possible that a strategy neither dominates nor is dominated by another strategy.

Assumption 1: A principled player who has internalized norm Ψ will not play an *equilibrium* strategy that is Ψ-dominated by another *equilibrium* strategy.

This assumption, which is common knowledge, might be justified on two grounds. First, it is arguably an intuitive way to model focal points –ultimately, though, this is largely an empirical matter. According to this idea, an equilibrium in which principled players follow binding norms is more obvious, and this coordinates players' beliefs.

Second, this refinement makes the model more precise and simplifies much the analysis. When searching for the equilibria of a game, one can first focus attention on those strategy profiles such that principled players respect internalized norms at *any* information set –note that they usually form a reduced set in many games. If any such profile is an equilibrium profile then it is trivially not norm-dominated by any other. Consequently, finding all those equilibria would finish the equilibrium analysis –if there is no such equilibrium, one should continue considering all profiles such that principled players respect the norm at all information sets except (maybe) one, and so on.

3. What Norm?

One may think of infinite correspondences satisfying definition 1. To obtain precise behavioral predictions in games and test the model, however, one must assume something

about the specific norms that principled players have internalized. Further, the number of norms should not be too high in order to keep the model tractable. Ideally, one norm should be able to explain a significant fraction of the experimental results.

This seems a difficult task because we know from sociologists and anthropologists' reports that human societies have myriads of norms, and it is not easy to discern which the key ones are. A prominent candidate, though, appear to be *norms of distributive justice* because concepts like fairness or justice are often employed to justify behavior.

These are prescriptions based on orderings of money distributions or, more generally, material welfare allocations. As an example, consider a norm that selects any action pointing towards an efficient and maximin outcome, *conditional on others doing the same*. More formally, let $X(\Delta)$ denote the set of all *monetary* allocations of lab game Δ.

Definition 3: Allocation $x=\{x_1,...,x_n\}\in X(\Delta)$ is efficient-cum-maximin (EM) if it maximizes function

$$F(x) = \sum_{i\in N} x_i + \delta \min_{i\in N} \{x_i\} \qquad (1)$$

over $X(\Delta)$, where $0<\delta$. An EM-path of Δ is a path leading to an EM-allocation of Δ. An EM-action is an action that belongs to an EM-path.

Definition 4 (the EM-Norm): If h is on one EM-path, the EM-norm selects only the EM-actions in $A(h)$. Otherwise, the EM-norm selects the whole set $A(h)$.

In other words: As far as everybody respects the EM-norm, then one must strive to achieve an EM-allocation; but if it is known *for certain* that at least one player has deviated then any behavior is allowed –i.e., the norm is conditional in an extreme form. The reader can verify that this is truly a norm –i.e., a *nonempty* correspondence selecting at least one action at any information set of any lab game. Similar norms of distributive justice are easily obtained by conveniently changing function (1). Thus, a norm embodying egalitarian concerns might correspond to function

$$F^{e}(x) = \min_{i\in N} \{x_i\} - \max_{i\in N} \{x_i\} . \qquad (2)$$

The EM-norm is extremely simple, and one can think of more sophisticated norms – for examples, consult López-Pérez (2005). In spite of that, I will assume in what follows that the EM-norm is the *only* norm that principled players care about. This simplifies much the analysis, and it is enough to replicate a good deal of the experimental facts.

An important reason why the EM-norm succeeds in explaining the evidence is because it assumes that people view both efficiency *and* maximin as basic ingredients of distributive justice. This implies that principled people are willing to sacrifice some money in order to promote efficiency and the welfare of the worst-off player(s), but not to

promote payoff equality. This point is particularly well illustrated and supported by the evidence coming from individual decision lab problems with externalities.[9]

I provide two examples. First, to show that people are willing to spend money for the sake of efficiency and maximin, consider a situation in which agent B has no say whereas player A must choose between (A, B) pecuniary allocations (4, 4) and (4-ε, 10). If ε and δ are small enough (more precisely, $\varepsilon \cdot (1+\delta) < 6$), the only EM-allocation is (4-ε, 10) and hence the EM-norm commends to choose it. My model predicts that behavior if player A is a principled one who has internalized the EM-norm and her internalization parameter γ is larger than ε -incidentally, she would clearly opt for (4, 4) if she were selfish. In contrast, she would unequivocally choose (4, 4) if she had internalized an egalitarian norm like that of (2).

Second, to show that people are not willing to spend money simply to promote equality, assume now that A must select either (3, 3) or (4, 6). According to my model A will always go for the latter allocation (whatever her type). In contrast, she would choose (3,3) if she had internalized an egalitarian norm and her γ was large enough.

In any case, more experiments are required to investigate what distributions people deem fair or just. For instance, it might be that nations or groups of people differ in what they view as fair. Thus, economists might be more concerned about efficiency than others - consult Fehr *et al.* (forthcoming) for evidence on this. Nevertheless, it must be pointed out that the model here is flexible enough to include such ideas. For instance, one could introduce some heterogeneity by assuming that some principled people have internalized the EM-norm while others have internalized an egalitarian norm.

4. Explaining the Evidence from the Lab

This section studies how the EM-norm affects cooperation, competition, coordination and punishment in several games. For simplicity, I assume that each player's type is common knowledge although I give some intuitions for the incomplete information case. In what follows, let ρ denote the fraction of principled players in the population.

4.1 Cooperation

The Prisoner's Dilemma Lab Game

To analyze the effects of the EM-norm on cooperation, it is convenient to consider first a *lab game* that has received huge attention from experimentalists: The Prisoner's Dilemma (PD) of Figure 1.

The two players (John and Ana in the example) *simultaneously* decide whether they cooperate (action C) or defect (action D). Both earn c monetary units if they cooperate, and d if both defect. Further, a unilateral defector gets a 'temptation' payment of t while

[9] Consult Frohlich and Oppenheimer (1992), Charness and Rabin (2002), Konow (2003), and Engelmann and Strobel (2004) for evidence, and López-Pérez (2004) for a more extensive discussion of this point.

a unilateral cooperator gets a normalized payoff of zero. Payoffs satisfy $t > c > d > 0$ - i.e., defection strictly dominates cooperation in *monetary* terms- and $2c > t$ so that (c,c) is the only EM-allocation and cooperation is the only EM-action. In short, there exists a stark conflict between self-interest and compliance with the norm.

		John: C	John: D
Ana	C	c , c	0 , t
	D	t , 0	d , d

Figure 1: (Ana's, John's) Monetary Payoffs in the PD Lab Game

To illustrate players' utility payoffs, assume that Ana is selfish and John is principled (other cases can be analogously analyzed). Trivially, Ana's utility coincides with her own pecuniary payoff. On the other hand, John gets some disutility (shame) if he deviates *unilaterally* from the EM-norm or if Ana does so (anger), but he feels no disutility if both players defect. Figure 2 represents all this.

Behavioral predictions are straightforward. First, mutual defection is the only Nash equilibrium if at least one player is selfish or if $\gamma < t - c$ and both players are principled. Second, mutual cooperation is the unique *refined* equilibrium if both players are principled and $\gamma \geq t - c$ - although mutual defection is also a Nash equilibrium, it can be ruled out because it is EM-dominated by mutual cooperation (assumption 1).

		John: C	John: D
Ana	C	c , c	$0, t - \gamma$
	D	t , $- \alpha \cdot t$	d , d

Figure 2: Utility Payoffs if Ana is Selfish and John is Principled

To sum up, *principled players are conditional cooperators*: They only cooperate if the other player is expected to cooperate as well. Intuitively, this idea also extends to a setting where players' types are private information. In that case, a principled player cooperates only if she believes with enough probability that her co-player is principled – i.e., the type of people who cooperate. More precisely, one can easily show that principled players cooperate in the *simultaneous* PD if priors are above threshold[10]

$$\rho^{sim} = \frac{d + \alpha \cdot t}{d + \alpha \cdot t - t + c + \gamma}. \qquad (3)$$

Consistent with the model, numerous experiments with one-shot prisoner's dilemmas –consult Rapoport and Chammah (1965), and Rabin (1993) for surveys; and Sally (1995) for a meta-analysis- find that a significant proportion of players cooperate, and that cooperation strongly depends on the expectation that the co-player will cooperate

[10] Note that condition $1 \geq \rho^{sim}$ requires $\gamma \geq t - c$.

as well. Thus, in one of the experimental treatments reported by Croson (2000), subjects played ten times a PD lab game against different co-players and they were asked to guess at the start of each round her co-player's future choice. 83% of the participants that guessed their counterpart would cooperate cooperated themselves. On the contrary, when participants expected that their opponent would defect, only 32% of them cooperated.

To finish, inspection of threshold (3) indicates that ρ^{sim} depends negatively on c and positively on t and d. The same occurs with the expected price of cooperation, that is, the *net*, expected material gain from defection. Taking as well into account that cooperation is hindered as ρ^{sim} grows, *a law of demand* follows: Cooperation decreases when its price increases. This prediction is again consistent with experimental evidence –see Rapoport and Chammah (1965, pp. 36-39), and Clark and Sefton (2001).

Fostering Cooperation: Sequential vs. Simultaneous Mechanisms

Assume now that the Prisoner's Dilemma is played in a sequential manner –e.g., Ana chooses after observing John's move. Apparently, this is a minor change. Indeed, the standard model predicts zero cooperation here, as in the simultaneous PD. If the second player is principled, though, the sequential mechanism changes players' incentives to comply with the EM-norm, and fosters cooperation.

To understand this point, note first that the sequential PD has a unique EM-path. In it, both players cooperate one after the other, hence reaching the EM-allocation. As a result, the EM-norm commends the first mover to cooperate. Further, it also commends the second mover to cooperate if the first mover cooperates, but allows *any action* if the first mover defects (definition 4). Consequently, the first mover is the *only* deviator from the EM-norm –i.e., the only person who 'misbehaves'- if both players choose defection. This is a subtle but key difference with the simultaneous PD, in which *both* players count as deviators if they mutually defect.

Given these norm prescriptions, the sequential PD has a unique Subgame Perfect Equilibrium for each parameter calibration –proving this is easy. I start by describing the second mover's equilibrium strategy. On one hand, a selfish second mover always defects. On the other hand, a principled second mover reciprocates the first mover's choice if she is principled and $\gamma \geq t - c$ -that is, she cooperates if he cooperated and defects if he defected- whereas she always defects if $\gamma < t - c$.

Experimental evidence from Hayashi et al. (1999) and Clark and Sefton (2001) is consistent with these predictions. Second movers often cooperate conditional on the first mover's choice, while unconditional cooperation is negligible. In addition, Clark and Sefton (2001) show that reciprocation falls as its material cost rises, something that is also consistent with my model, as reciprocation only occurs if $\gamma \geq t - c$.

Finally, the first mover's equilibrium strategy depends on his type and the second mover's. A selfish first mover cooperates only if the second mover is principled and $\gamma \geq t - c$ –this follows simply from $c > d$. A principled first mover only cooperates if the

second mover is principled and $\gamma \geq \min\{\alpha \cdot t + d, t - c\}$, *and* if the second mover is *selfish* and $\gamma \geq \alpha \cdot t + d$. In this latter case, the first mover cooperates even when he knows that his opponent will later defect. In that way, he avoids being the person who 'spoiled' cooperation, something that he finds particularly painful if $\gamma \geq \alpha \cdot t + d$.

The above mentioned results can be easily extended to an incomplete information setting. Since principled *second movers* reciprocate (if $\gamma \geq t - c$) and selfish ones always defect, a principled first mover cooperates in the *sequential* PD if $d - \gamma < (1 - \rho) \cdot (-\alpha \cdot t) + \rho \cdot c$, that is, if his prior is above threshold

$$\rho^{seq} = \frac{d + \alpha \cdot t - \gamma}{\gamma \cdot t + c}. \qquad (4)$$

Comparison between equations (3) and (4) indicates that $\rho^{sim} > \rho^{seq}$ if $\gamma \geq t - c$. Further, we have seen that selfish players never cooperate in the simultaneous PD but cooperate in the sequential PD if they move first and their prior is large enough. As a result, first movers' rate of cooperation in the sequential PD is significantly larger than the average cooperation rate in the simultaneous PD. This is consistent with the experimental evidence reported by Hayashi et al. (1999) and Clark and Sefton (2001).

To sum up, *moving first* in a sequential dilemma makes people more cooperative than if they choose simultaneously. This occurs for two key reasons. First, deviating from the EM-norm (or from any conditional norm of cooperation) in the sequential PD is *unilateral*. As a result, deviating is psychologically more disturbing than in the simultaneous PD, where *simultaneous* deviations are also possible. Second, selfish *first* movers with large enough priors find profitable to comply with the EM-norm because they understand that they can 'emotionally force' a principled second mover to comply as well, and so get more money than if both defect.[11]

On Positive Reciprocity

In some models of reciprocity -Rabin (1993), Levine (1998), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006)- one may distinguish between positive reciprocity (being kind with those who are kind) and negative one (being unkind with those who are unkind). Positive reciprocity implies that people are more kind with an active *and kind* player than with a passive player who makes no choice in the game.

To illustrate this, consider again the *sequential* PD lab game but assume now that the first mover –that is, John- has only action C available –i.e., he is a passive player. The only active player is Ana, who must choose therefore between (Ana's, John's) allocations (c, c) and $(t, 0)$. Clearly, the above mentioned reciprocity models predict that Ana will choose $(t, 0)$ significantly *more* if John is passive (call this the passive *cooperation* case) than if John is active and chose 'kind' action C (active *cooperation* case).

[11] See Rabin (1993, p. 1296) on this regard.

However, the available experimental evidence does not seem to support this prediction. Thus, Camerer (2003, pp.89-90) survey some results in this regard and concludes that the effect of positive reciprocity is insignificant or small.

What does my model predict in the passive and active cooperation settings? Contrary to other reciprocity models, it predicts *invariance,* or *no positive reciprocity*. In effect, Ana makes the same move in both cases *whatever* her type: She defects and attains allocation $(t,0)$ if she is selfish, and cooperates (the EM-action) if she is principled and $\gamma \geq t - c$. Hence, the model is more consistent with the experimental evidence.

The intuition behind the invariance result is twofold. On one hand, selfish types only care about available outcomes, and not about previous history, so that invariance makes no surprise. On the other hand, and in case nobody has broken the norm, it makes no difference for a principled player whether compliance happened because everybody was active and compliant or because everybody was passive –passive players, recall, respect the norm by definition. As a result, *principled players treat equally well both passive players and active compliant players*.[12]

What explains Invariance?

The previous comparison has pointed out one key difference between my model and other models of reciprocity. However, models like Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and the quasi-maximin model of Charness and Rabin (2002) also predict invariance. This occurs because these models assume that players only care about the distribution of income –i.e., players have consequentialistic utility functions. Hence, one can wonder whether the invariance result is the effect of subjects having consequentialistic preferences.

Although I will investigate this issue in more detail and give some evidence when studying punishment, it may be worth to consider again the sequential PD. Now, however, I will consider Ana's behavior in the following two situations: (i) John is active and has chosen action D (active *defection* case), and (ii) John is passive and has only action D available (passive *defection* case).

Since Ana faces (Ana's, John's) allocations (d, d) and $(0,t)$ in both cases, a consequentialistic model predicts invariance –i.e., Ana always chooses the same allocation. My model, on the contrary, predicts some variance if Ana is principled. On one hand, she chooses (d, d) in the active defection case because then she feels angry at John. On the other hand, Ana does not feel any anger at a passive John and moreover the EM-allocation is $(0,t)$ if $(2+\delta)\cdot d < t$. Consequently, she chooses $(0,t)$ if t and γ are large enough.

[12] Incidentally, it is often argued that rewards –i.e., acts that generate positive externalities on someone who has previously behaved in an approved way- are evidence in favor of positive reciprocity. The discussion here suggests that this assertion must be treated with care: Rewards might be alternatively explained as a result of people following social norms of distributive justice.

To sum up, *while a principled player may sacrifice some money to treat kindly a passive player, she will never do that with a deviator*. This idea is absent in a consequentialistic model, but it is important to appreciate why institutions have incentives to signal that they had no other choice when they made a tough decision that affected others. In that way, other agents will not perceive that choice as a violation of prevailing norms and hence will not get angry. For instance, many European governments and politicians who advocate for reforms in their Welfare States often argue that Globalization leaves them no way out. Though some of them may sincerely believe that, such type of arguments might be also part of a strategy designed to prevent voters' indignation.

Lab Games with n Players: Public Goods

In a simple Voluntary Contribution Mechanism (VCM) public good *lab game*, $n \geq 2$ subjects, each one with an endowment of e monetary units, choose *simultaneously* whether to contribute e to a public good or to keep the endowment for them.[13] Subject i's monetary payoff at terminal node z is given by $m \cdot e \cdot c(z)$ if she contributes and by $e + m \cdot e \cdot c(z)$ if she does not contribute, where m denotes the monetary payoff per unit of public good and is such that $m < 1 < n \cdot m$, and $c(z)$ stands for the number of players that contribute to the public good in the history of z. Since $m < 1$, the dominant strategy in material terms is not to contribute. Nevertheless, many experiments report aggregate contribution levels around 40-60% -for a survey, consult Ledyard (1995).

To get behavioral predictions, note first that the EM-norm commends every player to contribute because $1 < n \cdot m$. Let then n_p ($0 \leq n_p \leq n$) denote the number of principled players in the group (recall that I assume that players' types are common knowledge). For any n_p and γ, α, the VCM lab game has a unique *refined* equilibrium:

- If $\gamma < \alpha \cdot m \cdot e$, no player contributes.
- If $\gamma \geq \alpha \cdot m \cdot e$, no selfish player contributes while a principled player contributes only if $n_p = n$ or if $n_p < n$ and

$$e \cdot m \cdot n_p - \alpha \cdot e \cdot (1 + m \cdot n_p) \geq e + (m \cdot e - \gamma) \cdot (n_p - 1) \Leftrightarrow n_p \geq \frac{e \cdot (1 - m + \alpha) + \gamma}{\gamma - \alpha \cdot m \cdot e} = n_p^*(m, \alpha, \gamma). \quad (5)$$

The intuition is clear: Principled players respect the EM-norm if sufficiently many others do it as well. Note that there exist other equilibria if $n_p \geq n_p^*$, but they are EM-dominated because at least one principled player does not contribute in them and hence deviates from the EM-norm.

Observe also that n_p^*, the minimal number of principled agents necessary to sustain positive contributions (the *critical mass*), does not depend on the total number of

[13] In more complex VCM games, players are allowed to contribute a fraction of the endowment, and not only the whole one. This is unsubstantial for my model –I come back to this in the conclusion.

players n. Consequently, the probability that a group of n agents independently drawn from the population contains n_p^* or more principled players grows with n, and thus cooperation gets facilitated as n increases, a result supported by experimental evidence -see Isaac and Walker (1994).

The cooperative equilibrium has a natural counterpart if player's types are private knowledge. In that case, one can show that principled types contribute if their prior about ρ -i.e., the probability that a player is principled- is large enough. Hence, there exist a positive correlation between the expectations of a principled agent about aggregate contribution levels and her decision to contribute. Abundant experimental evidence bears this point –see Orbell and Dawes (1991), and Sonnemans et al. (1999).[14]

Experimental evidence -see Isaac and Walker (1988), and Ledyard (1995) for a survey- also shows that contribution levels raise if m increases. In this regard, inspection of equation (5) points out that n_p^* depends negatively on m only if γ and α are large and small enough, respectively, so that an increase in m will foster contributions only in those cases. In effect, if principled agents are very aggressive –i.e., α is high- and m is large, then anger costs rise substantially when they contribute –because contributing has the side-effect of increasing deviators' earnings. To find contribution optimal in that case, therefore, the emotional cost of defecting must be high enough.

4.2 Competition: Market Lab Games

Experimental evidence from a broad class of market lab games supports the standard prediction that prices *converge* to the competitive equilibrium –see, for instance, the survey in Fehr and Schmidt (1999, p. 829). Can a model of social norms explain that result? To study this point, consider a market game with proposer competition: $n-1$ sellers (proposers) make simultaneous price offers p_1, p_2, …, and p_{n-1} to sell one unit of a good to a single buyer (responder) who demands only one unit of the good. The buyer can accept the offer *she prefers* or reject all of them.

Assume that the responder values one unit of the good in V monetary units. Hence, the responder's monetary payoff if she accepts price offer p_i ($i \in \{1,2,..,n-1\}$) is $V - p_i$, whereas seller i's income is p_i -unsuccessful sellers get zero money. Finally, all players get no money if the responder accepts no offer.

Before applying the model of social norms to this game, it is convenient to consider first the standard prediction when all players are selfish. For any $n \geq 3$, the game has then

[14] In experiments with finitely repeated public goods games, aggregate contributions fall over time, getting very close to the zero level. The model here suggests that such phenomenon might be due to learning about the number of principled players. According to this, (some) principled subjects have upwardly biased priors that they revise when they observe actual contribution levels. This revision downwards might explain the decrease in contributions.

a basically unique SPE: The responder always accepts the minimum price offer and at least two proposers offer a price equal to zero.[15] The intuition why this equilibrium is unique is similar to that behind the Bertrand Duopoly equilibrium, and the reader is directed to a Microeconomics textbook for a proof. Finally, note that the standard equilibrium result is radically different if $n=2$ because then the proposer reaps the whole surplus V -this is the so-called *ultimatum game*; I briefly study it in section 4.4.

What does my model predict if $n \geq 3$? The key point here is that *all* allocations in this game are EM -the only exception are those allocations in which all players get zero, that is, those that follow a rejection by the responder. In effect, all allocations are efficient and moreover the worst pecuniary payoff is zero in all of them –if $n \geq 3$ there is always at least one unsuccessful seller who gets nothing. This implies that any price offer is an EM-action and that acceptance of any offer, but not rejection, is also an EM-action. Consequently, the utility payoffs of any type of player coincide with monetary ones except if the responder deviates from the norm, that is, rejects –then she suffers a utility cost if she is principled whereas principled sellers anger at her. It then follows that the game has a basically unique SPE that coincides with the standard one previously mentioned. Clearly, this result does not depend on players' types being common knowledge.

Some other models of social preferences *roughly* predict the same result. The model by Fehr and Schmidt (1999), for instance, predicts the standard solution if one assumes that the responder is restricted to accept or reject the *highest price* offer –note that this would not affect my model's predictions. On the contrary, a responder with high aversion to advantageous inequity would rather accept an egalitarian sharing of the surplus if she was given the opportunity to choose *any* offer –as I assumed in my analysis.

I finish this section on competition by briefly studying a market lab game with responder competition. Opposite to the game with proposer competition, this game has just one seller (proposer) and $n-1$ buyers (responders). The proposer moves first by proposing a selling price p and then each responder decides, unaware of other responders' choices, whether she accepts or reject p. All players receive a monetary payoff of zero if *all* responders reject p. In turn, the proposer gets p and the buyer $V-p$ if at least one responder accepts - a random draw selects with equal probability one of the accepting responders in case more than one accepts-, and all other responders receive zero.

Note that the standard model predicts a unique SPE. In it, responders accept any selling price while the proposer makes a price offer of $p=V$, thus reaping the whole surplus. Experimental evidence roughly supports this prediction –see Fehr and Schmidt (1999, p. 832) for references. Interestingly, my model also shares this unique prediction if $n \geq 3$ -the game is the ultimatum game if $n=2$, and subsection 4.4 studies it.

[15] Many strategy profiles satisfy this, but they only differ in the distribution of offers of the remaining $n-3$ sellers, which is inconsequential for the final result. Hence, the equilibrium outcome is unique.

The reasons are now familiar: All *Pareto-efficient* allocations in this game are EM-ones so that accepting any price offer is consistent with the EM-norm. Further, as rejection is never *pecuniary* profitable for principled or selfish responders, it follows that responders always accept in equilibrium, and a seller consequently asks for the whole surplus. This result is not affected if players' types are private knowledge, but it is nonetheless rather sensitive to the specific form of function (1). In case such function included an *additional* concern for inequality like that of function (2) the predictions might be different.

4.3 Coordination Lab Games and the Battle of the Sexes

Norms act as focal points, and hence it is natural to ask whether they increase coordination. To start with a simple example, consider a *lab* game in which two players choose simultaneously between raising a red flag or a blue one. (Row player's, Column player's) *monetary* payoffs are represented in Figure 3, and satisfy $a > b > 0$. For instance, each player gets a monetary units if both raise a red flag.

		Column Player: Red	Column Player: Blue
Row Player	Red	a , a	0 , 0
	Blue	0 , 0	b , b

Figure 3: Monetary Payoffs in a Coordination Game.

This is a simple coordination game. Real-life examples include deciding where and when we meet somebody, driving on the left/right side of a road, organizing teamwork and division of labor, or selecting an industry standard.[16]

The EM-norm clearly commends both players to raise the red flag. Hence, a principled player gets $-\gamma$ utils if she raises the blue flag unilaterally, 0 utils if the co-player alone raises the blue flag, and a or b if both raise the red or the blue flags, respectively. This implies that, whatever the players' types, there are two equilibria in pure strategies –i.e., {red, red} and {blue, blue}- and one in mixed strategies. Observe, however, that the {red, red} equilibrium EM-dominates all others.

Therefore, and if *at least one* player is principled, the model unequivocally predicts that the players will attain the efficient outcome (a, a). On the contrary, no equilibrium can be refined if both players are selfish: The model is then undetermined. Nevertheless, if it is assumed that the players' types are private information then it is intuitive that a selfish type will move red if the probability that the co-player is principled is large enough –more precisely, if $\rho > b/(a+b)$ holds, as one can easily prove.[17]

[16] For more examples of coordination games, consult Camerer (2003, 338-9).

[17] Observe that pre-play communication is not necessary for players to achieve coordination here. This suggests three conditions ensuring *efficient* play in a two-player game *without communication*: (a) It must be common knowledge (or highly likely) that one of the players has internalized the EM-

Other models are less precise. The standard *homo economicus* model and any model of social preferences or reciprocity, for instance, do not ensure that players will achieve the surplus-maximizing outcome. Of course, one might solve this by applying an equilibrium refinement, as my model does. The question is which one.

For instance, Harsany and Selten (1988) propose two major criteria to refine equilibria. One is risk-dominance, to which I will refer later, and the other one is *payoff dominance* or *Pareto efficiency*. An equilibrium is Pareto efficient if its associated vector of equilibrium utility payoffs is not Pareto dominated by any other equilibrium vector. Thus, equilibrium {red, red} in the game of Figure 3 is Pareto efficient, but {blue, blue} is not (assuming that utility coincides with money earned). The fact that {red, red} is the obvious solution in this game suggests that Pareto efficiency is the right refinement in this game. Nevertheless, refinements based on material efficiency –i.e., on the sum of equilibrium material payoffs- or on efficiency-cum-maximin work here as well as Pareto efficiency. Further, and as I will show later when analyzing other games like the Stag Hunt lab game, there is evidence contrary to the Pareto efficiency criterion.

The Battle of the Sexes

Figure 4 displays the matrix of (row player's, column player's) *monetary* payoffs in the 'Battle of the Sexes' lab game, where $a > b > 0$. Players move simultaneously.

	Boxing	Opera
Boxing	a , b	0 , 0
Opera	0 , 0	b , a

Figure 4: Monetary Payments in the Battle of the Sexes Lab Game.

Since both (a, b) and (b , a) are EM-allocations, the EM-norm allows both players to choose *any* action. Hence, utility payoffs coincide with monetary payments for *any* type of player –as in the standard model! This implies in turn that, for any matching of types, there exist two Nash equilibria in pure strategies, that is, {Boxing, Boxing} and {Opera, Opera}, and one equilibrium in mixed strategies -in it, the row player chooses Boxing with probability $a/_{a+b}$, while the column player chooses Boxing with probability $b/_{a+b}$. Note well that no strategy is EM-dominated by another one. [18]

Cooper *et al.* (1989) and Straub (1995) report experimental evidence on this lab game. Without pre-play communication, subjects fail to coordinate on a pure strategy

---

norm, (b) the norm must select a unique action profile (and for that there must be a unique EM-allocation), and (c) it must be *materially* profitable to follow that profile if the co-player does as well.

[18] In an interesting variation of the game, one player moves at time t whereas the other moves at time t+1 *without being informed of the other player's choice*. Standard theory (and my model) predicts here the same equilibria as in the simultaneous version. Nevertheless, a norm of 'first come, first served' might allow players to coordinate and reach the first mover's preferred *EM-allocation*.

equilibria. Although rather erratic, behavior appears to match the mixed strategy equilibrium prediction as subjects choose their preferred move with the highest probability.

The Stag Hunt Lab Game

Figure 5 depicts (row player's, column player's) *monetary* payoffs in the Stag Hunt lab game, where $a > b > 0$. Play is simultaneous and not repeated. The EM-outcome ensues if both players choose R, but that move is 'risky' in that one gets zero money if the other player fails to play R as well. On the contrary, action S is 'safe' because it gives a sure positive payoff of b.

	R	S
R	a , a	0 , b
S	b , 0	b , b

Figure 5: Monetary Payoffs in the Stag Hunt Lab Game

As the EM-norm selects action R, it follows that {R, R} is the unique *refined* equilibrium if at least one player is principled. Remaining equilibria –i.e., {S, S} and a mixed strategy equilibrium- are EM-dominated by {R, R}. In contrast, no equilibrium can be refined if both players are selfish and then there exist three equilibria: {S, S}, {R, R}, and a mixed strategy equilibrium in which both players choose R with probability b/a.

These results can be somehow extended to a setting where players' types are private information. Intuitively, a selfish player unequivocally plays R if she believes with enough probability that the co-player is principled –one can prove that, interestingly, the corresponding threshold prior *decreases* as difference $(a-b)$ increases. Other equilibria exist if the prior is low so that the rate of choice of action R is then undetermined.

In other words, efficient play is undetermined if $(a-b)$ sufficiently low. Experimental data apparently supports this indetermination. On one hand, Cooper et al. (1992) report results from one treatment in which each subject played the Stag Hunt lab game twenty times against different opponents. Payoffs (b =800; a =1000) were given in points that determined the probability of the player winning a lottery where winning players received \$1 and losing players received \$0. Thus, the difference $(a-b)$ was arguably small. The authors only give data from the last eleven periods, which show that play of the {S, S} equilibrium was prevalent -Clark et al. (2001) replicate these results.

On the other hand, Duffy and Feltovich (2002) use a payoff calibration and binary lottery procedure similar to those of Cooper et al. (1992), and show much larger levels of efficient play. They also report a large variance in the three sessions that they ran: In two of them, the frequency of efficient play is close to 50%, while in the other one it is 81%.

Experimental evidence from Straub (1995) is also consistent with my prediction that efficient play subtly depends on $(a-b)$. Alternatively, that data also supports the *risk*

dominance selection criterion proposed by Harsany and Selten (1988) –indeed, Straub (1995) suggest that interpretation.[19]

4.4 Punishment

Agent A punishes B when she imposes a cost on B without getting any immediate material reward as a result. According to my model, A punishes B only if B has transgressed a norm that A cares about and which A herself has not violated. This occurs because B's deviation triggers an aggressive emotion in A that goes associated with an impulse to retaliate.

To illustrate this with an example, consider the lab game tree at Figure 6, where only *monetary* payoffs are depicted. The first mover can offer either (player 1's, player 2's) allocation (8, 2) or (5, 5), and then the second mover can accept (A) or reject (R) the offer. In case she rejects it, both players get zero money. Otherwise, the offer is implemented. This lab game is a simplified version of an Ultimatum Game with stakes equal to 10 monetary units –the difference is that the range of offers in the ultimatum game consists of *all* possible divisions of the stakes. I stick to this simple version because it is sufficient to show how punishment works -for a detailed analysis of the model's predictions in the Ultimatum lab game, consult López-Pérez, 2004.

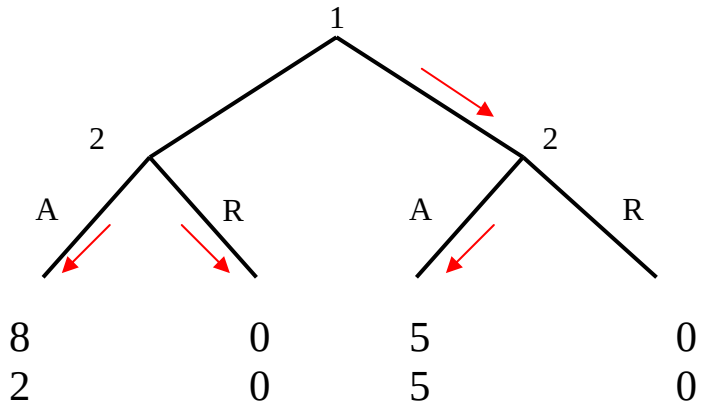

Figure 6: A Mini-Ultimatum Lab Game

As (5, 5) is the unique EM-allocation of this game, the EM-norm clearly commends player 1 to offer (5, 5) and player 2 to accept it. On the other hand, if player 1 deviates from the norm and offers (8, 2), the EM-norm allows player 2 to choose *any* move. Arrows in Figure 6 indicate that the associated action is selected by the EM-norm.

The game has essentially a unique SPE. In it, a selfish second mover accepts any offer. Further, a principled second mover accepts offer (5, 5), rejects (8, 2) if $0 > 2 - 8 \cdot \alpha$ and accepts it if $0 < 2 - 8 \cdot \alpha$. In the marginal case $\alpha = 0.25$, a principled second mover is indifferent between accepting or rejecting (8, 2) and there are two SPE then.

[19] An equilibrium is risk dominant if it maximizes the product of the gains from unilateral deviation. If both Stag Hunt players are selfish and $2b > a$, equilibrium {S, S} is the only risk-dominant one because $(b-0)\cdot(b-0) > (a-b)\cdot(a-b)$.

In turn, the first mover's offer depends on both players' types, as Figure 7 indicates. The first column in this matrix shows player 1's type, while the first row shows player 2's type. For instance, player 1 abides by the EM-norm and offers (5, 5) independently of the co-player's type if she is principled and $\gamma \geq 3$.[20]

Player 2's type is… / Player 1's type is…	… selfish or principled with $\alpha < 0.25$.	… principled with $\alpha > 0.25$.
… selfish or principled with $\gamma < 3$.	(8, 2)	(5, 5)
… principled and $\gamma \geq 3$.	(5, 5)	(5, 5)

Figure 7: Player 1's SPE offer depending on her type and the second mover's.

Note that this result can be easily extended to an incomplete information setting. The only caveat in that case is that a selfish first mover or a principled one with $\gamma < 3$ would condition her choice on her prior about the second mover's type -the reader may easily compute the minimal prior that makes an offer of (8, 2) optimal.

Experimental data from ultimatum games –see Camerer (2003, pp. 48-55) for an informative survey- confirms that the 50-50 offer is almost always accepted, whereas low offers face a high probability of rejection. Studies also show that "very large changes in stakes have only a modest effect on rejections",[21] something that is barely consistent with my model –if a principled second mover rejects (8, 2) in the lab game of Figure 6 then she also rejects offer (8·k, 2·k) when stakes are k>0 times bigger.

The model shows that punishment depends on parameter α -which, incidentally, could be estimated from experimental data. In fact, if one assumed that principled players are heterogeneous regarding their aggressiveness –i.e., parameter α-, *a law of demand* would follow: The more costly punishment is the less of it there is. To see this, consider a slightly modified version of the lab game at Figure 6 in which allocation (6, 4) replaces allocation (8, 2). Since (5, 5) is still the only EM-allocation, a principled second mover will anger if she is offered (6, 4). Nevertheless, punishing (i.e., rejecting) such offer is more costly than rejecting offer (8, 2) and hence only optimal if α is relatively large –more precisely, if $\alpha > 2/3$ holds. To sum up, principled agents use relatively costly punishment technologies only if they are aggressive enough.

Another interesting issue concerns *responsibility* (or 'intentions', to use a usual terminology). Experimental evidence indicates that responsibility is crucial to understand *who is punished* (Blount 1995), and the model is consistent with this because it predicts that only wrongdoers get punished. To understand some implications of this, assume that player 1 has no say in the lab game of Figure 6 and that his move is decided by a random

[20] There are two SPE if $\gamma = 3$ and the second mover accepts (8, 2). However, the equilibrium in which player 1 offers (8, 2) is EM-dominated and can thus be ruled out.

[21] Camerer (2003, pp. 61).

device –thus, player 1 is not to be blamed for anything that happens in the game. As the EM-norm does not restrict non-human choices, a principled second mover will not anger at any offer and hence will not reject it. Therefore, and in comparison with the intentional treatment, the model predicts a *smaller* rate of rejection in the random treatment, something that is consistent with the results reported by Blount (1995).

The word 'intentions' also refers sometimes to the *influence of non-chosen alternatives*. To illustrate this point, consider a slight variation of the lab game of Figure 6, in which allocation (10, 0) replaces allocation (5, 5). Compare now the rejection rate of offer (8, 2) in this new game and in the former game. Does the model predict a difference? The answer is yes. As offer (8, 2) constitutes a deviation from the EM-norm when the alternative is (5, 5), but not when the alternative is (10, 0), the model of social norms clearly predicts a larger rejection rate in the former case –in fact, the model predicts that nobody rejects (8, 2) if the alternative is (10, 0). In general, as a norm may select different actions depending on the available alternatives, an act may constitute misbehavior and hence be punished in one game but not in another, *even* if it has the same material consequences in both cases. This is highly consistent with the experimental evidence –see Camerer (2003, p. 81-82).

5. Comparison with Other Utility Models

It can be illustrative to compare the behavioral predictions of my model with those from other models.[22] With regard first to cooperation and punishment, the model has been shown to be consistent with seven well-replicated experimental phenomena:

(1) A significant proportion of people cooperate in a *simultaneous* PD lab game, or contribute in a one-shot public good lab game.

(2) Subjects also contribute in a *sequential* PD, and the rate of first movers' cooperation is larger than average cooperation in the simultaneous PD.

(3) Subjects give money to passive players (dummies).

(4) Subjects tend to treat equally kindly both dummies and *kind* active players (absence of positive reciprocity).

(5) Many subjects sacrifice equality of payments in order to increase efficiency and/or the worst-off player's payoff.

(6) Punishment depends on the whole menu of alternatives, not only on the available ones.

(7) Subjects do not punish dummies (responsibility).

Figure 8 indicates whether other utility theories are consistent with facts (1) to (7). Entry YES indicates that the corresponding theory is consistent with the fact, whereas entry NO indicates the opposite. For brevity, I consider just four models, each one representing a different research line in the existing literature. Models of inequity aversion

[22] Consult López-Pérez (2004) for a more lengthy discussion that includes the analysis of other games like Ultimatum, Dictator, Trust, Best-Shot, and Cournot duopoly games.

like Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) represent pure consequentialistic models in which people only have distributional concerns –other examples include the model of quasi-maximin preferences of Charness and Rabin (2002). Rabin (1993) is a pure reciprocity model with no distributional concerns, as Dufwenberg and Kirchsteiger (2004). These two models are based on the Psychological Game Theory of Geanakoplos *et al.* (1989), as Falk and Fischbacher (2006). However, the latter introduces both reciprocal and distributional concerns. Finally, Levine (1998) is a model of type-based reciprocity.

Facts / Theories	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rabin (1993)	YES	NO	NO	NO	NO	YES	YES
Levine (1998)	YES	YES	YES	NO	YES	YES	NO
Inequity aversion	YES	YES	YES	YES	NO	NO	NO
F&F (2006)	NO	YES	YES	NO	NO	YES	NO

Figure 8: Predictions by Other Utility Models

The interested reader is directed to the relevant papers for a detailed explanation of these predictions. However, I would like to remark that both mutual cooperation and defection constitute equilibria in the simultaneous PD (or in a one-shot public good game) if both players are sufficiently averse to advantageous inequality. This implies that models of inequity aversion are undetermined with regard to the cooperation level in the PD game.[23] The same occurs with Rabin (1993), which can be only applied to two-player, normal form games, and Dufwenberg and Kirchsteiger (2004).

Most models can explain the experimental results from market lab games. With regard to the battle of the sexes, predictions by inequity aversion models or Falk and Fischbacher (2006) are highly dependent on the value of the parameters. For instance, Fehr and Schmidt (1999) predict equilibria {Boxing, Opera} and {Opera, Boxing} if both players are very inequity averse. Rabin (1993), and Dufwenberg and Kirchsteiger (2004) share those predictions if both players have a "strong enough emotional reaction to each other's behavior" (Rabin, 1993, p. 1285). In this regard, it might be interesting to investigate a sequential version of the game to test these theories, which predict that some second movers will intentionally mismatch. Finally, models of inequity aversion predict in the Stag Hunt game the same pure strategy equilibria as the standard model, and hence zero is the lowest possible level of coordination on the efficient outcome.

6. Conclusion and Extensions

[23] If one assumes, as Fehr and Schmidt (1999, page 842) suggest, that efficiency and symmetry act as a focal point then the cooperative equilibrium is focal, and the predictions by Fehr and Schmidt (1999) for this game would be similar to mine. However, this focal point is not consistent with other lab evidence, as the analysis of the Stag Hunt game previously showed.

This paper offers a model of social norms and shows that a large set of experimental evidence can be explained if one assumes that (some) people care about a particular norm of distributive justice. The model explains not only why people cooperate in some cases and compete in others, or how punishment work, but also why norms may sometimes enhance coordination on efficient outcomes. The model appears to be empirically more relevant than other models of non-selfish preferences. Moreover, it is much simpler and precise than other models of reciprocity.

There are some possible ways to extend the model. A natural one is considering other norms that the EM-norm. For instance, one could assume that some people have internalized a norm of honesty, and study how it affects communication. This is an interesting issue and it seems that the model here has a comparative advantage when compared with other models of social preferences.

One could also think of more realistic norms of distributive justice. The EM-norm is too strict in that it allows any behavior after one deviation occurs. Less draconian norms would select in that case only those actions leading towards an allocation that maximizes the material welfare of those who *have hitherto respected* the norm - López-Pérez (2005) gives particular examples. Further, the EM-norm is probably too austere in that it focuses on EM-actions. However, people seem to have a more flexible view of what is correct: 'Small' deviations from the ideal moral behavior –e.g. the EM-path in this model- are usually considered valid as well, and they do not trigger anger.

Some of the motivational hypothesis of the model could be also relaxed. For instance, the model assumes that bad feelings do not depend on the specific deviation one makes from an internalized norm. But it seems realistic to assume that remorse is higher depending on the material consequences of the deviation[24] –e.g., cheating in a medical article should generate more remorse than cheating in an economics paper! This hypothesis and an additional one that the marginal utility of money is decreasing would explain, for instance, why participants in public good games often contribute something between zero and their endowment.

As a final remark, the model here should motivate further experimental research on social norms, emotions, and reciprocity. Further, it might be used to study how norms and emotions affect bargaining, collusion between firms, conflict, charity giving, revolutions, team behavior, and voting, to give some examples.

[24] I have investigated this point in López-Pérez (2005).

Bibliography

- Arrow, Kenneth J. (1974). *The Limits of Organization*. Norton & Company.
- Becker, G. (1996). *Accounting for Tastes*, Harvard University Press.
- Blount, S. (1995). "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences", Organizational Behavior & Human Decision Processes 63(2), 131–144.
- Bolton, G. E., and A. Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition", American Economic Review, 90(1), pp. 166-93.
- Camerer, C. (2003). *Behavioral Game Theory-Experiments in Strategic Interaction*, Princeton University Press.
- Charness, G., and M. Rabin (2002). "Understanding Social Preferences with Simple Tests", Quarterly Journal of Economics, 117, 817-869.
- Clark, K., S. Kay, and M. Sefton (2001). "When are Nash Equilibria Self-Enforcing? An Experimental Analysis", International Journal of Game Theory 29, 495-515.
- Cooper, R., D. DeJong, R. Forsythe, and T. Ross (1989). "Communication in the Battle of the Sexes Game: Some Experimental Results", RAND Journal of Economics 20(4), 568-587.
- Cooper, R., D. DeJong, R. Forsythe and T. Ross (1992). "Communication in Coordination Games", Quarterly Journal of Economics 107, 739-771.
- Croson, R. T. A. (2000). "Thinking like a Game Theorist: Factors affecting the Frequency of Equilibrium Play." Journal of Economic Behavior and Organization, 41, 299-314.
- Duffy, J., and N. Feltovich (2002). "Do Actions Speak Louder Than Words? Observation vs. Cheap Talk as Coordination Devices", Games and Economic Behavior, 39, 1-27.
- Dufwenberg, M., and G. Kirchsteiger (2004). "A Theory of Sequential Reciprocity", Games and Economic Behavior, 47, 268-98.
- Elster, J. (1989). "Social Norms and Economic Theory", Journal of Economic Perspectives, 3(4), 99-117.
- Engelmann, D., and M. Strobel (2004). "Inequality Aversion, Efficiency and Maximin Preferences in Simple Distribution Experiments", American Economic Review, 94(4), 857-869.
- Falk, A., and U. Fischbacher (2006). "A Theory of Reciprocity", Games and Economic Behavior 54, 293-315.
- Fehr, E. and K. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation", Quarterly Journal of Economics, 114(3), 817-68.
- Fehr, E., and K. Schmidt (2002). "Theories of Fairness and Reciprocity - Evidence and Economic Applications", in M. Dewatripont, L. Hansen and St. Turnovsky

(Eds.), *Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs*, Cambridge University Press.

- Fehr, E., M. Näf, and K. Schmidt. "The Role of Equality, Efficiency, and Rawlsian Motives in Social Preferences: A Reply to Engelmann and Strobel." Forthcoming in American Economic Review.
- Frohlich, N., and J. A. Oppenheimer (1992). *Choosing Justice: An Experimental Approach to Ethical Theory*. Berkeley and LA: University of California Press.
- Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). "Psychological Games and Sequential Rationality." Games and Economic Behavior 1, 60-79.
- Gintis, H. (2003). "The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms." Journal of Theoretical Biology 220(4), 407-418.
- Harsany, J. C., and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*, MIT Press.
- Hayashi, N., E. Ostrom, J. Walker, and T. Yamagishi (1999). "Reciprocity, Trust and the Sense of Control: A Cross-Societal Study." Rationality and Society, 11, 27-46.
- Isaac, R. M., and J. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism." Quarterly Journal of Economics, 103, pp. 179-99.
- Isaac, R. M., J. Walker, and A. Williams (1994). "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing very Large Groups." Journal of Public Economics, 54, pp. 1-36.
- Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). "Fairness and the Assumptions of Economics." Journal of Business, 59 (4, 2), 285-300.
- Konow, J. (2003). "Which Is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature*, 41 (4), pp. 1186-1237.
- Ledyard, J. (1995). "Public Goods: A Survey of Experimental Research", in J. Kagel and A. E. Roth (Eds.), *Handbook of Experimental Economics*, Princeton Univ. Press.
- Levine, D. K. (1998). "Modeling Altruism and Spitefulness in Experiments." Review of Economic Dynamics, 1, 593-622.
- López-Pérez, R. (2004). "Emotions Enforce Fairness Norms", mimeo.
- López-Pérez, R. (2005). "Guilt and Shame in Games", mimeo.
- Orbell, J., and R. Dawes (1991). "A Cognitive Miser' Theory of Cooperators' Advantage." American Political Science Review, 85, 515-28.
- Parsons, T. (1967). *Sociological Theory and Modern Society*, New York: Free Press.
- Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics", American Economic Review 83, 1281-1302.
- Rapoport, A. and A. M. Chammah (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, MI: University of Michigan Press.
- Schelling, T. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.

- Sonnemans, J., A. Schram, and T. Offerman, 1999. "Strategic Behavior in Public Good Games: When Partners Drift Apart." Economics Letters 62, 35-41.
- Straub, P. G. (1995). “Risk Dominance and Coordination Failures in Static Games”, The Quarterly Review of Economics and Finance 35 (4), 339-363.
- Sugden, R. (1989). “Spontaneous Order”, Journal of Economic Perspectives, 3(4), 85-97.